Combining heterogeneous spatial datasets with process-based spatial fusion models: A unifying framework

Wang, Craig ; Furrer, Reinhard

Abstract: In modern spatial statistics, the structure of data has become more heterogeneous. Depending on the types of spatial data, different modeling strategies are used. For example, kriging approaches for geostatistical data; Gaussian Markov random field models for lattice data; or log Gaussian Cox process models for point-pattern data. Despite these different modeling choices, the nature of underlying data-generating (latent) processes is often the same, which can be represented by some continuous spatial surfaces. A unifying framework is introduced for process-based multivariate spatial fusion models. The framework can jointly analyze all three aforementioned types of spatial data or any combinations thereof. Moreover, the framework accommodates different likelihoods for geostatistical and lattice data. It is shown that some established approaches, such as linear models of coregionalization, can be viewed as special cases of the proposed framework. A flexible and scalable implementation using R-INLA is provided. Simulation studies confirm that the prediction of latent processes improves as one moves from univariate spatial models to multivariate spatial fusion models. The framework is illustrated via a case study using datasets from a cross-sectional study linked with a national cohort in Switzerland. The differences in underlying spatial risks between respiratory disease and lung cancer are examined in the case study.

DOI: https://doi.org/10.1016/j.csda.2021.107240

Computational Statistics and Data Analysis 161 (2021) 107240

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Combining heterogeneous spatial datasets with process-based spatial fusion models: A unifying framework

Craig Wang [a,*], Reinhard Furrer [a,b], for the SNC Study Group

[a] Department of Mathematics, University of Zurich, Switzerland
[b] Department of Computational Science, University of Zurich, Switzerland

ARTICLE INFO

Article history:
Received 19 July 2020
Received in revised form 23 March 2021
Accepted 25 March 2021
Available online 20 April 2021

Keywords:
Bayesian methods
Change of support problem
Data fusion
Gaussian process

ABSTRACT

In modern spatial statistics, the structure of data has become more heterogeneous. Depending on the types of spatial data, different modeling strategies are used. For example, kriging approaches for geostatistical data; Gaussian Markov random field models for lattice data; or log Gaussian Cox process models for point-pattern data. Despite these different modeling choices, the nature of underlying data-generating (latent) processes is often the same, which can be represented by some continuous spatial surfaces. A unifying framework is introduced for process-based multivariate spatial fusion models. The framework can jointly analyze all three aforementioned types of spatial data or any combinations thereof. Moreover, the framework accommodates different likelihoods for geostatistical and lattice data. It is shown that some established approaches, such as linear models of coregionalization, can be viewed as special cases of the proposed framework. A flexible and scalable implementation using R-INLA is provided. Simulation studies confirm that the prediction of latent processes improves as one moves from univariate spatial models to multivariate spatial fusion models. The framework is illustrated via a case study using datasets from a cross-sectional study linked with a national cohort in Switzerland. The differences in underlying spatial risks between respiratory disease and lung cancer are examined in the case study.

1. Introduction

In modern spatial statistics, researchers are dealing with increased heterogeneity in the structure of collected spatial data. Different data sources may contain overlapping information concerning the same research questions. In addition, combining different datasets can mitigate the problem of limited sample size and imperfect measurements (Diggle and Lophaven, 2006). In public health and epidemiology, joint analysis of multiple diseases borrows information from related diseases to account for the underlying correlations and to uncover common risk factors.

Spatial models are useful when residuals exhibit correlation in space after known covariates are accounted for in a regression-type setting. Spatial data has long been classified into three categories, namely geostatistical (point-level) data, lattice (area-level) data and point-pattern data (Cressie, 1991). Depending on data types, different statistical models that capture the residual spatial correlation are used. In a nutshell, (1) geostatistical data are observations with geo-coordinates. For example, rainfall measurements at locations of weather stations (Kyriakidis et al., 2001). The strength of dependency can be modeled as a function of the distance separation between two locations. (2) Lattice data can be

* Corresponding author.Correspondence to: Department of Mathematics, University of Zurich, Winterthurerstrasse 190, CH-8057, Switzerland.
E-mail address: craig.wang@math.uzh.ch (C. Wang).

https://doi.org/10.1016/j.csda.2021.107240

either gridded or irregularly aligned, and occur in the form of aggregated observation over areas. They are often collected in epidemiology, such as disease prevalence of each district (Chammartin et al., 2016). Gaussian Markov random field (GMRF) models such as conditionally autoregressive models are typically used to capture the spatial dependency between neighboring areas. Finally, (3) point-pattern data are observations where the locations themselves are stochastic. They are used to model epidemiological data of case locations (Gatrell et al., 1996) or other event locations such as epicenter of earthquakes (Ogata, 1988). One approach to model such data is using a point process, where the intensity function may depend on observed covariates (Gelfand et al., 2010).

Despite there are different modeling strategies for different types of data, a common purpose of all the aforementioned statistical models is to capture the residual spatial dependency between different observations after known covariates are accounted for. A natural way to do this is using Gaussian processes to model continuous spatial surfaces, which represent the underlying scientific process that drives response variables together with observed covariates. It is beneficial to jointly model heterogeneous types of spatial data on the underlying process instead of observed spatial support. Several authors (Moraga et al., 2017; Shi and Kang, 2017; Wang et al., 2018) demonstrated improved prediction performance when multiple datasets were modeled together compared to a single dataset via both simulation study and real-world applications. In addition, better scientific interpretations can be obtained by modeling different data types at the underlying process (Møller et al., 1998; Kelsall and Wakefield, 2002).

In order to model spatial data at a different support than that from the observed data, change of support problem occurs. There are methods proposed to analyze lattice data and point-pattern data using Gaussian process-based models. Kelsall and Wakefield (2002) modeled aggregated disease counts using a Gaussian process approach. Point-pattern data can be linked to Gaussian process with a log-Gaussian Cox process (LGCP) (Møller et al., 1998; Diggle et al., 2013). Taking a step further, several approaches used Gaussian process as a basis to fuse spatial data of different types. We refer to these approaches as spatial fusion models. Some earlier work on spatial fusion models (Berrocal et al., 2010; McMillan et al., 2010; Sahu et al., 2010) implemented efficient algorithms for specific model structures bundled with their data applications. Moraga et al. (2017) proposed a model to analyze spatial data available in both geostatistical and lattice type, with the same set of covariates and a response variable observed at two different spatial resolutions. Wilson and Wakefield (2018) extended the work by allowing non-Gaussian response variables. Shi and Kang (2017) proposed a fixed rank kriging-based fusion model to combine multiple lattice-type remote sensing datasets. Wang et al. (2018) utilized a nearest neighbor Gaussian process to combine geostatistical and lattice data with a flexible model structure.

In general, spatial fusion models are challenged by the trade-off between flexibility and computational efficiency, i.e., more flexible modeling structures require more generalizable inferential tools that come with a higher computational cost. For example, the fusion model in Wilson and Wakefield (2018) with normally-distributed response variables can take advantage of stochastic partial differential equation (SPDE) approaches and INLA which took few minutes to run. A more flexible modeling structure for Poisson-distributed response requires using Hamiltonian Monte Carlo-based inference method and took several weeks.

In this paper, we build on the work in Wang et al. (2018) and extend fusion modeling in two aspects. In terms of flexibility, our framework incorporates an additional data type, namely point-pattern data. To the best of our knowledge, this is the first spatial fusion framework that incorporates all three types of spatial data. We additionally allow arbitrary combinations of those three data types in multivariate settings. We propose a unifying framework that includes features of several well-established models, such as linear models of coregionalization (LMC) (Wackernagel, 2003; MacNab, 2016), spatial factor model (Wang and Wall, 2003) and shared component model (Knorr-Held and Best, 2001). In terms of computational cost, we offer an efficient SPDE and INLA-based implementation which significantly reduces computation time from hours to minutes for thousands of observations. Last but not least, we benchmark the performance of our implementation in terms of prediction and parameter estimation in simulation studies.

The rest of this paper is structured as follows: Section 2 introduces the unifying framework and explicitly shows its link to existing spatial models. Section 3 discusses implementation strategy using the SPDE approach and INLA. Section 4 illustrates the framework using two simulated scenarios and a re-analysis on epidemiological datasets to model respiratory disease risk in Kanton Zurich, Switzerland. Finally, we end with a summary followed by discussions on identifiability problems and research outlook in Section 5.

2. Process-based spatial fusion model

2.1. The unifying framework

For $j = 1, \ldots, \ell$, we let $\boldsymbol{Y}_j = \{Y_{j,1}, \ldots, Y_{j,n_j}\}$ denote the jth response variable with n_j observations, with a likelihood that belongs to the exponential family. Each of the ℓ responses can take any of the following data types: (i) geostatistical data, observed at locations $\boldsymbol{s}_j \in D \subseteq \mathcal{R}^2$; (ii) lattice data observed at areas $\boldsymbol{a}_j \subset D$; or (iii) point-pattern data that has been discretized to regular fine grid containing mostly zeros or ones, observed at gridded locations $\boldsymbol{v}_j \in D$, where $\boldsymbol{Y}_j$ denotes the number of events in the grid cell containing $\boldsymbol{v}_j$. Further, we let $\boldsymbol{X}_j$ denote a full (column) rank $n_j \times p_j$ matrix of spatially-referenced covariates that are observed at the same spatial units as the corresponding response variables, $\boldsymbol{\beta}_j$ denote a vector $p_j \times 1$ of fixed effect coefficients. We assume there is a $q \times 1$ vector of zero-mean, unit variance, independent latent Gaussian processes $\boldsymbol{w}$ having a $\ell \times q$ design matrix $\boldsymbol{Z}$ with rows $\boldsymbol{Z}_j$, i.e. $\boldsymbol{Z}_j\boldsymbol{w}$ is the jth linear combination of Gaussian

processes. Each Gaussian process is parameterized by its own covariance function. Finally, non-linear operator $B_j(\cdot)$ subsets and aggregates some components of $\boldsymbol{Z}_j\boldsymbol{w}$ such that the linear combination matches the spatial resolution of corresponding response variable. Overall, the framework is formulated as

$$g_j(\mathbb{E}[\boldsymbol{Y}_j \mid \boldsymbol{\beta}_j, \boldsymbol{Z}_j, \boldsymbol{w}]) = \boldsymbol{X}_j\boldsymbol{\beta}_j + B_j(\boldsymbol{Z}_j\boldsymbol{w}), \qquad j = 1, \ldots, \ell, \tag{1}$$

where $g_j(\cdot)$ is a link function that corresponds to the conditional distribution of $\boldsymbol{Y}_j$.

Although the latent processes have a continuous index, we work with a finite set of locations in practice. The set of locations $\mathcal{U}$ to be modeled in the latent processes $\boldsymbol{w}$ comprises of locations where geostatistical data are observed, locations of sampling points for lattice data and gridded locations for point-pattern data. The non-linear operator $B_j(\cdot)$ takes a different form depending on data types. For geostatistical and point-pattern data, the non-linear operator $B_j(\cdot)$ subsets $\boldsymbol{Z}_j\boldsymbol{w}$ to the corresponding locations of the jth response variable. For lattice data, a change of support problem arises (Gotway and Young, 2002) since we only observe aggregated information while the underlying process is continuous. When the link function is linear, $B_j(\cdot)$ subsets $\boldsymbol{Z}_j\boldsymbol{w}$ to the sampling point locations and aggregates them to the corresponding areas by taking averages. With non-linear link functions, $B_j(\cdot)$ first applies an inverse link function and then aggregates.

When modeling lattice data, we employ a sampling-points approximation approach to stochastic integrals (Gelfand et al., 2001; Fuentes and Raftery, 2005) for aggregating latent processes. Let $\boldsymbol{s}'$ denote the set of all sampling points and let H denote the number of sampling points of each area. Further, to simplify the notation, we denote with w an arbitrary component of the q spatial latent process. For the ith area $\boldsymbol{a}_{ji}$ in jth response, under a linear link function we obtain

$$w(\boldsymbol{a}_{ji}) = |\boldsymbol{a}_{ji}|^{-1} \int_{\boldsymbol{u}\in\boldsymbol{a}_{ji}} w(\boldsymbol{u})d\boldsymbol{u} \approx \frac{1}{H}\sum_{\boldsymbol{s}'\in\boldsymbol{a}_{ji}} w(\boldsymbol{s}'), \tag{2}$$

with $|\boldsymbol{a}_{ji}|$ being the area of $\boldsymbol{a}_{ji}$, i.e. the latent process at area $\boldsymbol{a}_{ji}$ is approximated by aggregating the process at sampling points within the area. Ecological bias will arise from non-linear link functions (Greenland, 1992), that is, the sum of non-linear functions applied to individual w is different from the non-linear function applied to the sum of individual w's. For a general link function $g_j(\cdot)$, we can avoid such bias by using the following approximation instead of (2),

$$w(\boldsymbol{a}_{ji}) = g_j\left(|\boldsymbol{a}_{ji}|^{-1} \int_{\boldsymbol{u}\in\boldsymbol{a}_{ji}} g_j^{-1}\left(w(\boldsymbol{u})\right) d\boldsymbol{u}\right) \approx g_j\left(\frac{1}{H}\sum_{\boldsymbol{s}'\in\boldsymbol{a}_{ji}} g_j^{-1}\left(w(\boldsymbol{s}')\right)\right). \tag{3}$$

Under the identity link function $g(x) = x$, Eq. (3) is equivalent to Eq. (2). For Poisson response with log link function $g(x) = \log(x)$, Eq. (3) becomes

$$w(\boldsymbol{a}_{ji}) = \log\left(|\boldsymbol{a}_{ji}|^{-1} \int_{\boldsymbol{u}\in\boldsymbol{a}_{ji}} \exp\left(w(\boldsymbol{u})\right) d\boldsymbol{u}\right) \approx \log\left(\frac{1}{H}\sum_{\boldsymbol{s}'\in\boldsymbol{a}_{ji}} \exp\left(w(\boldsymbol{s}')\right)\right). \tag{4}$$

Typically, a small H between 5 to 10 is chosen to balance the trade-off between computational efficiency and model accuracy (Fuentes and Raftery, 2005; Liu et al., 2011).

2.2. Linking to existing models

Our proposed unifying framework utilizes elements from existing literature and combines them to create a flexible yet efficient spatial fusion model framework. As a result, there are some connections in terms of model structure between this framework and other established methods in spatial statistics. At the same time, they share the same potential identifiability issues.

In univariate settings, the unifying framework allows us to model each type of spatial data individually with a latent Gaussian process. When we have geostatistical data, the framework results in a geostatistical regression (Cressie, 1991). With Poisson-distributed lattice data, we obtain a sampling-points approximation to the model used in Kelsall and Wakefield (2002), which is an alternative modeling strategy to Besag–York–Mollé model (Besag et al., 1991). With point-pattern data, we obtain a discretized LGCP (Møller et al., 1998).

In multivariate geostatistical data settings, the design matrix $\boldsymbol{Z}$ plays a pivotal role in the identifiability of model parameters. When the number of independent Gaussian processes is less than the number of responses $q < \ell$, we obtain a spatial factor model (Wang and Wall, 2003). The latent spatial factors are assumed to have zero-mean unit-variance Gaussian processes, such that $\boldsymbol{Z}$ controls the variance (partial sill) of latent processes. When $q = \ell$, we obtain a general coregionalization framework (Wackernagel, 2003; Schmidt and Gelfand, 2003). A similar LMC framework also exists for lattice data (MacNab, 2016). Identifiability issues occur in the LMC since the number of latent values to be estimated in the latent processes is equal to the total number of observations in response variables. Additional spatial hyper-parameters and fixed-effect coefficients also need to be estimated. For this reason, regularization is done via one of the following: (1) employing empirical Bayes method by fixing some of the hyper-parameters; (2) choosing informative prior distributions in Bayesian models; or (3) using a lower triangular matrix for $\boldsymbol{Z}$ (Schmidt and Gelfand, 2003). In cases of $q > \ell$, we acquire a similar model structure as shared component models (Knorr-Held and Best, 2001) for Gaussian processes,

where multiple outcomes have their own latent spatial components plus some shared spatial components. In this setting, the values in $\boldsymbol{Z}$ need to be even further constrained to avoid identifiability issues (Knorr-Held and Best, 2001).

Our framework is also linked to other process-based spatial data fusion models that combine geostatistical and lattice data types. When we let the response variables represent the same information with different data types, we obtain the model presented in Wilson and Wakefield (2018), where an explicit relationship is used to link multiple response variables. If we further allow different information to be represented in the response variables, we reach the generalized spatial fusion model framework proposed in Wang et al. (2018).

To the best of our knowledge, there is no existing approach or implementation that jointly models all three types of spatial data in a multivariate framework. With those links to the existing approaches, our framework extends upon them by combining different features and enhances the overall flexibility of spatial fusion models.

3. Model implementations

It is well known that fitting full Gaussian processes in Bayesian models is computationally expensive in both univariate and multivariate settings. Marginalized and conjugate Gaussian process models dramatically save computation time but they can only be used when geostatistical data with normally-distributed outcomes is fitted (Banerjee et al., 2014; Zhang et al., 2019). There are several approaches to reduce the computational burden, such as low rank (Cressie and Johannesson, 2008; Banerjee et al., 2008; Stein, 2008) and sparse (Furrer et al., 2006; Rue et al., 2009; Datta et al., 2016) methods. Some of those approaches are utilized in existing spatial fusion models. Shi and Kang (2017) adapted the spatial basis function approach from fixed rank kriging (Cressie and Johannesson, 2008). Moraga et al. (2017) used integrated nested Laplace approximations (INLA) (Rue et al., 2009). Wang et al. (2018) exploited the nearest neighbor Gaussian process (NNGP) (Datta et al., 2016). In this paper, we offer an efficient implementation strategy for the unifying spatial fusion model framework. The strategy follows Wilson and Wakefield (2018) to use the SPDE approach and INLA, with additional approximations for non-linear link functions.

Although the computational efficiency can be improved by using NNGPs instead of full Gaussian processes, it is still not feasible to fit multiple latent processes with more than thousands of locations in $\mathcal{U}$. Therefore, we choose the SPDE approach and INLA for the implementation of the fusion models.

Lindgren et al. (2011) established a connection between Matérn Gaussian process and GMRFs through a SPDE approach, where a Gaussian process over finite collection of locations can be approximated by triangulating the spatial domain and using a weighted sum of basis functions as

$$w_{\mathcal{U}} \approx \sum_{k=1}^{\mathcal{M}} r_k \phi_k, \tag{5}$$

where $\mathcal{M}$ is the number of points in the triangulation, r_k are Gaussian distributed weights and ϕ_k are basis functions. The weights $\boldsymbol{r} = [r_1, r_2, \ldots, r_{\mathcal{M}}]$ forms a GMRF which follows a multivariate normal distribution with a sparse precision matrix (Rue and Held, 2005) that makes computation efficient. In this approximation approach, the covariance function of the Gaussian process must be a member of the Matérn family defined as

$$C_{\boldsymbol{u}_i,\boldsymbol{u}_j} = \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)}(\sqrt{2\nu}\|\boldsymbol{u}_i - \boldsymbol{u}_j\|/\phi)^{\nu} K_{\nu}(\sqrt{2\nu}\|\boldsymbol{u}_i - \boldsymbol{u}_j\|/\phi), \tag{6}$$

where $\|\boldsymbol{u}_i - \boldsymbol{u}_j\|$ is the Euclidean distance between $\boldsymbol{u}_i$ and $\boldsymbol{u}_j$, K_{ν} is the modified Bessel function of second kind with integer order ν, σ^2 is the partial sill, ϕ relates to the spatial range and ν is the smoothness parameter. The approximation in Eq. (5) can be written as $w_{\mathcal{U}} \approx A\boldsymbol{r}$, where A is a projection matrix that maps a GMRF defined on the triangulation mesh nodes to the observations' locations.

INLA, which is suitable for a wide range of latent Gaussian models, can be used to fit this approximation approach for $\nu \in (0, 1]$ (Lindgren and Rue, 2015). The key to implementing the spatial fusion models in INLA lies within the projection matrix, with a different structure required for each data type (Krainski et al., 2018).

- For geostatistical data, the ith row of the projection matrix corresponds to the ith location, it is filled with zeros except where (1) the location is on the jth vertex, then the jth column is ones or (2) the location is within a triangulation area, then three cells get values based on a mixture of barycentric based weights from three neighboring vertices of the triangulation.
- For lattice data, we construct a projection matrix that links the ith area with the mean value of GRF at mesh nodes which falls into the ith area in analogous to Eq. (2). If the link function is linear, increasing the mesh density will increase the number of mesh nodes that fall into each area, therefore, better approximates the average. However, it is sufficient to have a sparse mesh for non-linear link functions (Follestad and Rue, 2003).
- Finally, for the point-pattern data, we use an augmentation approach by Simpson et al. (2016), which avoids discretizing the spatial domain into grid cells. The projection matrix is built as an identity matrix with a dimension equal to the total number of mesh nodes, row-binded with a projection matrix that is constructed on observed locations in the same way as the geostatistical case.

The final model fitting is done by stacking the projection matrices corresponding to each response variable together using `inla.stack()` and assigning appropriate priors using the **INLA** (Lindgren and Rue, 2015) package in R.

Recent advances in **INLA** (Martins et al., 2013) such as allowing multiple likelihoods and 'copy' feature made this implementation possible. When the response variables follow different distributions, the `family` argument in `inla()` is used to assign them. The underlying latent process needs to be specified only once and then assigned to different response variables using the `copy` argument in `f()` when specifying a model formula. Example R code is available in the supplementary materials.

4. Illustrations

In this section, we conduct two simulation studies and an analysis of epidemiological datasets to illustrate our proposed framework. We do not include comparisons with existing methods since they are either inflexible to be fitted with our model structure or computationally infeasible for the datasets. All results are obtained in R version 3.5.0 (R Core Team, 2018), on a Linux server with 256 GB of RAM and two Intel Xeon 6-core 2.5 GHz processors. All R code used in the simulation studies is provided in the supplementary materials.

4.1. Simulation study one

We are interested in modeling a single latent spatial process within a $[0, 10] \times [0, 10]$ square, using three spatial response variables with one response variable from each data type. First, we simulate a zero-mean GRF on densely uniformly distributed locations with a covariance matrix $C\left(\cdot, \cdot; \sigma^2, \phi\right)$. We then sub-sample 200 locations to obtain the latent process at observed locations. For lattice observations, we divide the square into 100 Voronoi cells and compute aggregated GRF from all locations while accounting for ecological bias using Eq. (3). In addition, we generate covariates X_1 and X_2 for geostatistical and lattice response by sampling independently from a standard normal distribution. Afterwards, we generate a normally-distributed geostatistical response at the same sampled locations and a Poisson-distributed lattice response for each area. For point-pattern observations, we simulate from the same GRF on a coarse 20×20 grid, then exponentiate the values to obtain intensity at the grid cells. Afterwards, we generate Poisson point process using each intensity value multiplied by cell area and an offset term as the final intensity. In summary, the response variables are generated according to

$$\begin{aligned} \boldsymbol{Y}_1 \mid \boldsymbol{\beta}_1, \boldsymbol{w}, \tau_1^2 &\sim \mathrm{N}\left(\boldsymbol{X}_1\boldsymbol{\beta}_1 + B_1(\boldsymbol{w}), \tau_1^2\boldsymbol{I}\right), \\ \boldsymbol{Y}_2 \mid \boldsymbol{\beta}_2, \boldsymbol{w} &\sim \mathrm{Pois}\left(\exp(\boldsymbol{X}_2\boldsymbol{\beta}_2 + B_2(\boldsymbol{w}))\right), \\ \boldsymbol{Y}_3 \mid \boldsymbol{w} &\sim \mathrm{Pois}\left(A\exp(B_3(\boldsymbol{w}))\right). \end{aligned} \tag{7}$$

In the simulation, we use an exponential covariance function ($\nu = 0.5$), i.e. $C(\boldsymbol{u}_i, \boldsymbol{u}_j; \sigma^2, \phi) = \sigma^2\exp(-\|\boldsymbol{u}_i - \boldsymbol{u}_j\|/\phi)$, $\boldsymbol{u}_i$, $\boldsymbol{u}_j$ two spatial locations in $\mathcal{R}^2$. The influence of different sample sizes and combination of spatial hyperparameters on predictive performance was investigated in a fusion model by Wang et al. (2018), therefore, we only consider a single setup by setting $\sigma^2 = 0.5$ and $\phi = 1$. In addition, we set $\boldsymbol{\beta}_1 = (1, 5)^\top$, $\boldsymbol{\beta}_2 = (1, 1.5)^\top$ and $\tau^2 = 1$. A is a constant term that controls the density of point process, and it is assigned as the product of grid cell area and an offset which takes value 0.25. $B_1(\cdot)$ and $B_3(\cdot)$ multiplies $\boldsymbol{w}$ with a matrix of zeros and ones to subset it to the locations of $\boldsymbol{Y}_1$ and gridded locations of $\boldsymbol{Y}_3$, while B_2 applies Eq. (4) on the subset of sampling locations for $\boldsymbol{Y}_2$.

We consider seven different model specifications within our proposed framework: three univariate models using a single data type each, namely one of geostatistical, lattice and point-pattern data; three fusion models using different combinations of two data types; and a multivariate fusion model combining all three response variables. We use penalized complexity (PC) prior for Matérn GRF (Fuglstad et al., 2019) with α fixed at 1.5, corresponding to the exponential covariance function. In addition, we choose the median practical spatial range to be 2 (corresponds to the median of ϕ being 1) and the probability of σ greater than 1.7 is 5%. The remaining priors are default options from R-INLA (Lindgren and Rue, 2015), i.e., a zero-mean normal distribution with precision 0.001 for the coefficients $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$; and a Gamma distribution with shape being 1 and rate being 0.00001 for the precision τ^2. The simulation is repeated 100 times, each model runs just under one minute.

We choose an additional 1600 sites from a regular grid to evaluate the predictive performance of models on the latent process in terms of posterior standard deviation and root mean squared prediction errors (RMSPE) given by

$$\mathrm{RMSPE} = \left(\frac{1}{1600}\sum_{i=1}^{1600}\left(w(\boldsymbol{u}_i) - \hat{w}(\boldsymbol{u}_i)\right)^2\right)^{1/2} \tag{8}$$

under each scenario. In Eq. (8), $\hat{w}(\boldsymbol{u}_i)$ denotes the posterior median. Note that the latent process is kept the same under each simulation for a fair comparison. The prediction sites are located at the centers of a 40×40 grid that uniformly covers the sampling domain. Their predictive performance is shown in Fig. 1. The first Venn diagram shows the average posterior standard deviation over 100 simulations under different models. The second Venn diagram shows average RMSPEs. The point-process data only model has the lowest posterior standard deviations, however, its RMSPE is the largest. Overall, the RMSPEs are smaller in multivariate fusion models compared to univariate process-based models. The joint modeling of all three types of spatial data has the lowest RMSPE on the prediction of the latent process at unobserved locations, demonstrating the benefits of spatial fusion models.

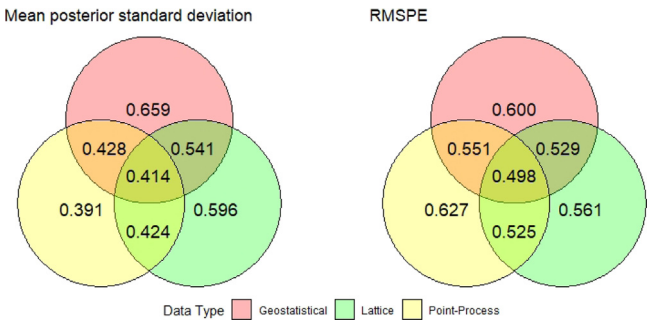

Fig. 1. Venn diagram of mean posterior standard deviation and root mean squared prediction error for the unifying fusion framework fitted to each data type (and combinations thereof). Values in overlapping areas indicate results from models with multiple data types.

4.2. Simulation study two

In the second simulation study, we simulate a new scenario with three response variables of different types and two latent processes to evaluate the parameter estimates of a multivariate fusion model. We firstly simulate two independent zero-mean unit-variance GRFs $\boldsymbol{w} = (w_1, w_2)^\top$ with exponential covariance function distributed on the spatial domain of $[0, 100] \times [0, 100]$ square, then compute the sub-sampled and aggregated GRF for each response variable in the same way as in simulation one.

Each response depends on the latent processes via the design matrix

$$\boldsymbol{Z} = \begin{bmatrix} 1.2 & 0 \\ 0.5 & 1.2 \\ 0 & 1 \end{bmatrix}. \tag{9}$$

The first geostatistical response variable only depends on the first latent process, the second lattice response variable depends on both latent processes, while the third point pattern response variable depends only on the second latent process. The response variables are generated as follows,

$$\begin{aligned} \boldsymbol{Y}_1 \mid \boldsymbol{\beta}_1, \boldsymbol{w}, \tau_1^2 &\sim \mathrm{N}\left(\boldsymbol{X}_1\boldsymbol{\beta}_1 + B_1(\boldsymbol{Z}_1\boldsymbol{w}), \tau_1^2\boldsymbol{I}\right), \\ \boldsymbol{Y}_2 \mid \boldsymbol{\beta}_2, \boldsymbol{w} &\sim \mathrm{Pois}\left(\exp(\boldsymbol{X}_2\boldsymbol{\beta}_2 + B_2(\boldsymbol{Z}_2\boldsymbol{w}))\right), \\ \boldsymbol{Y}_3 \mid \boldsymbol{w} &\sim \mathrm{Pois}\left(A\exp(B_3(\boldsymbol{Z}_3\boldsymbol{w}))\right), \end{aligned} \tag{10}$$

where $\boldsymbol{Y}_1$ consists of 500 geostatistical observations, $\boldsymbol{Y}_2$ has 100 lattice observations and $\boldsymbol{Y}_3$ represents the number of events observed at each of 400 cells on a 20×20 grid. In addition, we set $\boldsymbol{\beta}_1 = (3, 5)^\top$, $\boldsymbol{\beta}_2 = (0.5, 2)^\top$, $\phi_1 = 5$, $\phi_2 = 25$ and $\tau_1^2 = 0.5$. $B_1(\cdot)$, $B_2(\cdot)$ and $B_3(\cdot)$ are defined similarly as in simulation one. Since we have two latent processes in the simulation, using any of the univariate model or fusion model with two response variables can lead to identifiability problem. Therefore, we estimate the parameters using the unifying spatial fusion model with three responses only. The model and their prior specifications are the same as in simulation one, except for the spatial range parameter and the design matrix. The prior for both ϕ_1 and ϕ_2 is a PC prior with median practical spatial range 20 (corresponds to the median of ϕ being 10). For implementation purposes, the design matrix components are parameterized differently in INLA. Z_{11} and Z_{32} are treated as σ_1 and σ_2 of each latent process, while Z_{21}/Z_{11} and Z_{22}/Z_{32} are coefficients for the latent processes with variance σ_1^2 and σ_2^2. The former has the same prior as in simulation one, the latter has a normal prior with a mean of 1 and a standard deviation of 10. The simulation is run 100 times.

The parameter estimates based on posterior medians are displayed in Fig. 2. The PC prior in R-INLA penalizes complex structure in GRF hence tends to have a slightly over-estimated range and underestimated coefficients in $\boldsymbol{Z}$ (Fuglstad et al., 2019). One example of the posterior median of fitted latent processes at locations with geostatistical observations is shown in Fig. 3. The figure shows both w_1 and w_2 are fitted well and the plot of fitted w_1 and w_2 shows their independence structure assumed in the model. The root mean squared errors are 0.54 and 0.48 for the first and second latent processes. The computation time for the INLA implementation of the fusion model is 11 min.

4.3. Application to LuftiBus-SNC dataset

In spatial epidemiology, joint analysis of multiple diseases with similar etiology allows us to separate underlying risk factors into shared and disease-specific components. In this analysis, we examine the disease-specific spatial risk surface of lung cancer and shared spatial risk components between lung cancer and respiratory disease while taking PM_{10} (particulate matter with diameter $< 10\ \mu m$) pollution surface into consideration. Previously, a similar analysis was done in Wang et al. (2018) considering combined lung cancer and respiratory disease risk only.

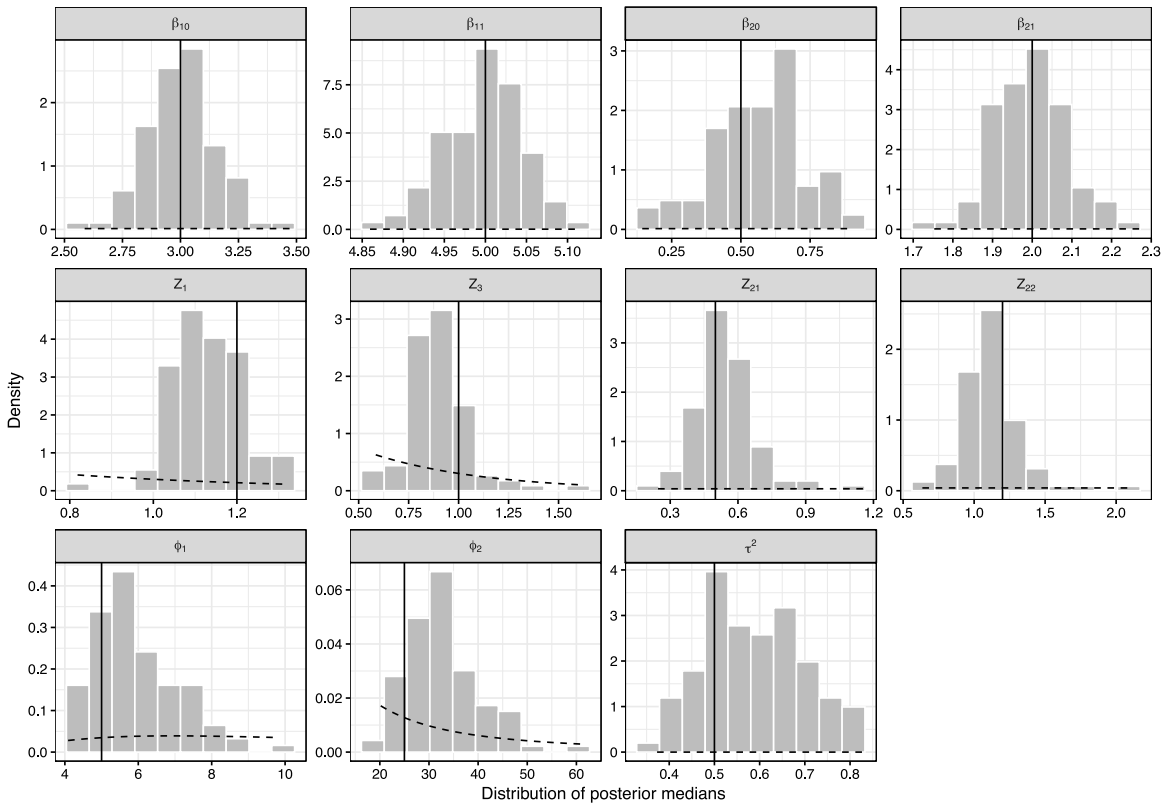

Fig. 2. Distribution of posterior median estimates from the unifying spatial fusion model in simulation two. The black vertical line indicates the true parameter value, dashed curves indicate prior density.

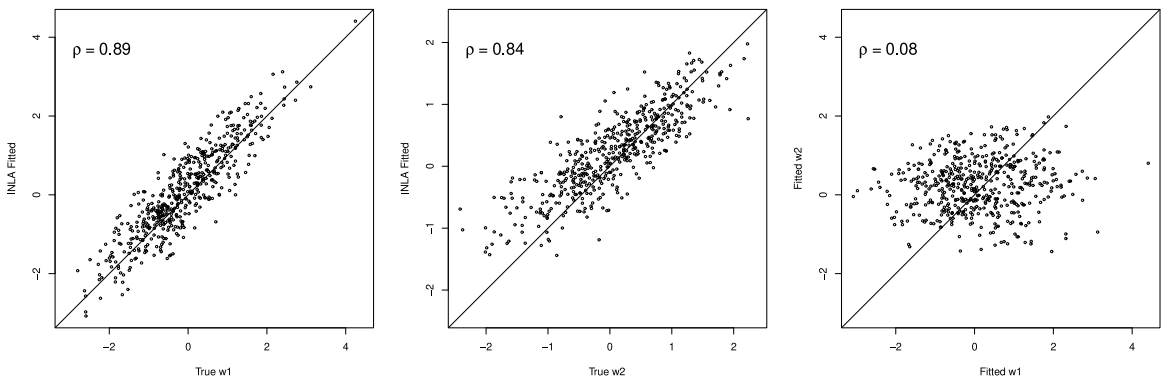

Fig. 3. One example of true versus fitted latent process at locations with geostatistical observation in simulation two, for w_1, w_2 and fitted w_1 against w_2 respectively. Pearson's correlation coefficients ρ are displayed.

Chronic lung disease contributes substantially to morbidity and mortality worldwide, with chronic obstructive pulmonary disease (COPD) being the third leading cause of death (Lozano et al., 2012). Forced expiratory volume in one second (FEV1) is a measure of the amount of air a person can exhale during a pulmonary test and it can be used to diagnose disease and predict respiratory-related mortality (Menezes et al., 2014). While respiratory disease and lung cancer share many common risk factors such as smoking and exposure to air pollution, it is of interest to examine the lung cancer-specific spatial risk component. It may provide insights into identifying risk factors that are solely associated with lung cancer.

Initiated as a health promotion campaign by Lunge Zürich (2017) in Switzerland, the 'LuftiBus' project collected lung function measurements including FEV1 and demographic information from local residents. Data from LuftiBus observed between 2003 and 2012 were deterministically linked with the census-based Swiss National Cohort (SNC)

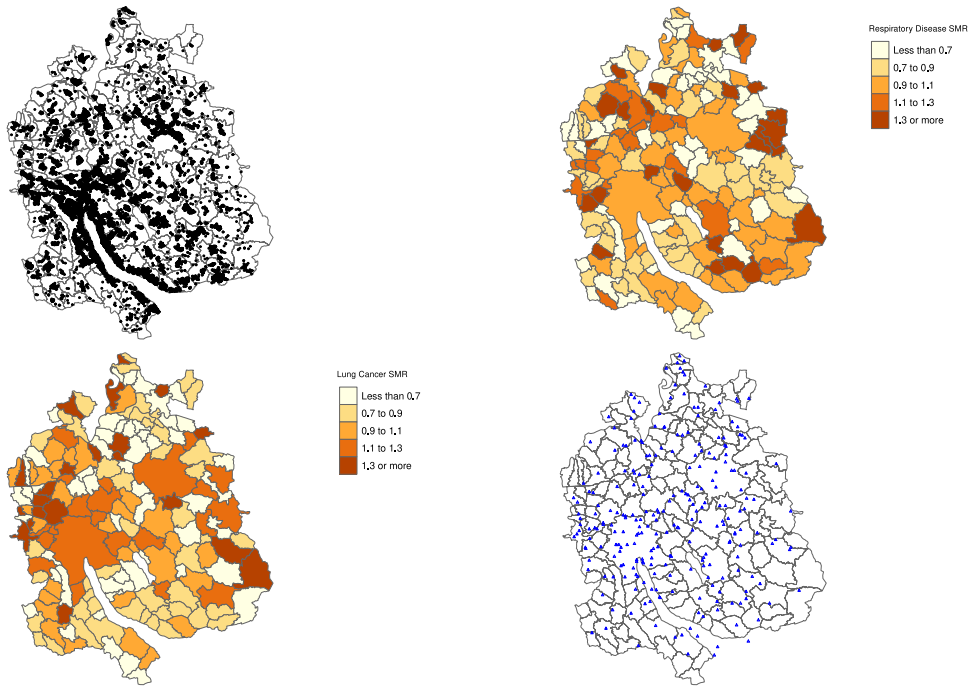

Fig. 4. Data used in the fusion model. Top left: locations of FEV1 observations. Top right: respiratory standardized mortality ratio. Bottom left: lung cancer standardized mortality ratio. Bottom right: sampled locations with density proportional to PM_{10} annual mean.

(Bopp et al., 2008). The purpose was to obtain 44,071 individuals with demographic, health and environmental variables in Switzerland. More importantly, the linkage provides us with the residential location of individual participants.

For lattice data, we use the computed expected cause-specific (respiratory and lung cancer, respectively) mortalities in each municipality, adjusted by 5-year age-group and gender based on the SNC data ($n = 572{,}993$). For point process data, we sample from a PM_{10} pollution map in 2010 (Geoinformation Kanton Zurich, 2015) to obtain proxy pollution sources. The probability of sampling each location is proportional to standardized PM_{10} annual mean value. We obtain 253 locations which are mainly distributed along highway roads.

We assume three latent spatial risk surfaces that are associated with respiratory disease and lung cancer. The first risk surface, $\boldsymbol{w}_1$, is shared between FEV1, respiratory mortality and lung cancer mortality, while the second risk surface $\boldsymbol{w}_2$ is lung cancer-specific risk surface and the third risk surface $\boldsymbol{w}_3$ is shared between PM_{10} and the other three disease-related variables. Typically with lattice data, multivariate conditional autoregressive models allow us to jointly analyze multiple responses and identify different latent components (Jin et al., 2005). Because municipal boundaries are artificial, we argue that a continuous spatial surface is a more natural modeling assumption. Therefore, we use our process-based unifying framework to conduct the analysis. Another advantage is that it allows us to incorporate the rich FEV1 data from Luftibus as well as additional pollution data.

The fusion model is structured as

$$\begin{aligned}
\boldsymbol{Y}_{\text{FEV1}} &\sim \mathrm{N}(\beta_{1,0} + \beta_{1,1}\,\boldsymbol{X}_{\text{age}} + \beta_{1,2}\,\boldsymbol{X}_{\text{gender}} - B_1(Z_{11}\,\boldsymbol{w}_1) - B_1(Z_{13}\,\boldsymbol{w}_3),\, \tau^2\boldsymbol{I}),\\
\boldsymbol{Y}_{\text{resp}} &\sim \mathrm{Pois}\big(\boldsymbol{E}_{\text{resp}} \exp(\beta_{2,0} + B_2(Z_{21}\,\boldsymbol{w}_1) + B_2(Z_{23}\,\boldsymbol{w}_3))\big),\\
\boldsymbol{Y}_{\text{cancer}} &\sim \mathrm{Pois}\big(\boldsymbol{E}_{\text{cancer}} \exp(\beta_{3,0} + B_2(Z_{31}\,\boldsymbol{w}_1) + B_2(Z_{32}\,\boldsymbol{w}_2) + B_2(Z_{33}\,\boldsymbol{w}_3))\big),\\
\boldsymbol{Y}_{\text{pm}} &\sim \mathrm{Pois}\big(A \exp(\beta_{4,0} + B_3(Z_{43}\boldsymbol{w}_3))\big),
\end{aligned} \tag{11}$$

where $\boldsymbol{E}_{\text{resp}}$ and $\boldsymbol{E}_{\text{cancer}}$ are the expected cause-specific mortalities. We set some coefficients of the design matrix $\boldsymbol{Z}$ to zero and directly depict the association between each response variable and the latent processes using the resulting product of $\boldsymbol{Z}$ and $\boldsymbol{w}$. The terms $B_1(Z_{11}\,\boldsymbol{w}_1)$ and $B_1(Z_{13}\,\boldsymbol{w}_3)$ are set to be negative since FEV1 is assumed to decrease as the latent risks increase. In addition, we add an additional intercept for $\boldsymbol{Y}_{\text{pm}}$ to indicate constant background pollution.

More than 60% of the FEV1 measurements in the linked dataset are located in the Canton of Zurich, therefore, we restrict our analysis to the Canton of Zurich. In addition, we focus the analysis on people who are 40 years or older, which results in 16,160 geostatistical observations. We use PC prior for the latent components with $\alpha = 1.5$ corresponding to exponential covariance function, median practical range of 1 km and median σ of 1. Fig. 4 shows the locations of geostatistical observation, standardized mortality ratio for respiratory disease and lung cancer and point process locations

Table 1
Parameter estimates and their 95% posterior credible intervals (95% CI) for the LuftiBus-SNC dataset. ϕ_1, ϕ_2 and ϕ_3 are in meters.

Parameter	Median	95% CI	Parameter	Median	95% CI
$\beta_{1,0}$	4.74	(4.70, 4.79)	z_{11}	0.0809	(0.0539, 0.122)
$\beta_{1,1}$	0.907	(0.891, 0.923)	z_{13}	0.0566	(0.0324, 0.0796)
$\beta_{1,2}$	−0.0375	(−0.0382, −0.0368)	z_{21}	0.0855	(0.0537, 0.137)
$\beta_{2,0}$	−0.0643	(−0.130, −0.00387)	z_{23}	0.141	(0.0868, 0.228)
$\beta_{3,0}$	−0.119	(−0.191, −0.0502)	z_{31}	0.0902	(0.0386, 0.172)
$\beta_{4,0}$	−0.0501	(−0.178, −0.0833)	z_{32}	0.500	(0.208, 1.170)
σ^2	0.268	(0.263, 0.274)	z_{33}	0.159	(0.0931, 0.251)
ϕ_1	367	(173, 690)	z_{43}	0.123	(0.0824, 0.168)
ϕ_2	307	(98, 1030)			
ϕ_3	1970	(1110, 3130)			

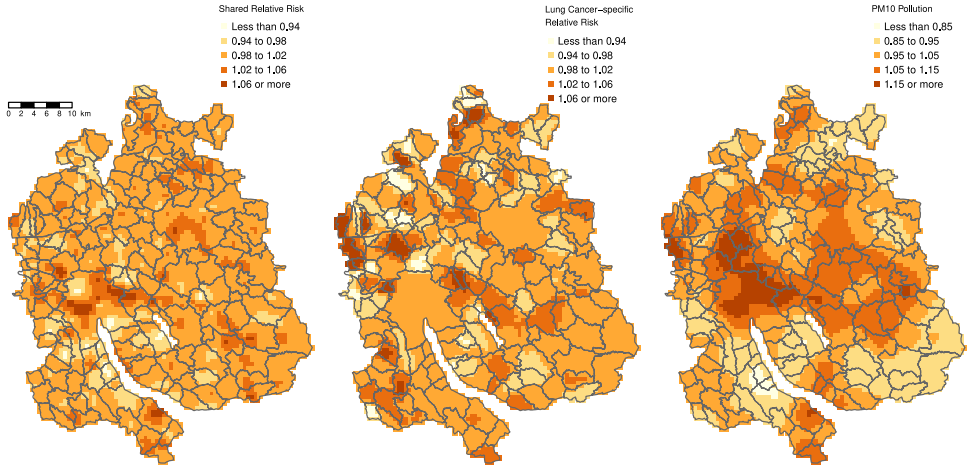

Fig. 5. Estimated spatial relative risk surfaces. Left: shared component between FEV1, respiratory mortality and lung cancer mortality, $\exp(\boldsymbol{w}_1)$. Middle: lung cancer mortality-specific component, $\exp(\boldsymbol{w}_2)$. Right: shared component between PM_{10} and all three disease-related response variables, $\exp(\boldsymbol{w}_3)$.

for pollution source proxy. Table 1 shows parameter estimates and Fig. 5 shows the transformed posterior estimates of the latent processes representing relative risk surfaces in the Canton of Zurich. The shared risk components between FEV1, respiratory mortality, and lung cancer mortality is the highest in urban areas, with an effective range of 1.1 km (95% CI: 0.5, 2.1) based on the exponential covariance function. The estimated relative risk is computed by exponentiating the latent process, which varies between 0.91 and 1.13. Meanwhile, high-risk areas of lung cancer-specific components are scattered around the Canton of Zurich, mainly in the north and west regions with an effective range of 0.9 km (95% CI: 0.3, 3.1). The variability is larger than the shared component with values between 0.90 and 1.20. The lung cancer-specific risk component is modeled via lattice data $\boldsymbol{Y}_{\text{cancer}}$ only, hence it appears to have some block-wise structures compared to the shared component and have larger variability in the range estimation. Such results complement classical lattice data-based disease mapping approaches to identify potential disease hotpots at a finer resolution. The pollution surface is shared between all four response variables and has the longest effective range with 5.9 km (95% CI: 3.3, 9.4). It is the strongest around city centers.

5. Summary and discussions

We have proposed a unifying process-based statistical framework to handle spatial data fusion. The framework allows simultaneously modeling all three types of spatial data, namely geostatistical, lattice and point-pattern data, to be easily incorporated into a single multivariate spatial model. This framework contains theoretical and computational elements from several existing literatures: the implementation uses the SPDE approach (Lindgren et al., 2011) and INLA (Rue et al., 2009), the sampling point approximation approach for modeling lattice data is adopted from Fuentes and Raftery (2005) and data augmentation (Simpson et al., 2016) is used in INLA. We have combined all of the individual elements and constructed this unifying framework. The framework extends upon existing flexible spatial fusion models (Wang et al., 2018; Wilson and Wakefield, 2018) by making point-pattern data also compatible, hence it completes all three spatial data types. We have shown in the first simulation study that it is advantageous to conduct multivariate analysis using multiple spatial datasets if they are available.

Identifiability issues arise when there is more than one latent spatial process in the fusion model. Similar concern has been brought up in other multivariate spatial models (Ren and Banerjee, 2013; Knorr-Held and Best, 2001). Since the model becomes invariant under certain orthogonal transformations, the design matrix $\boldsymbol{Z}$ is not identifiable. Knorr-Held and Best (2001) proposed a specific constraint on the relationship among the individual elements of the design matrix. Ren and Banerjee (2013) proposed to constrain one element of each row in the design matrix to be strictly positive and to have an ordered spatial range parameter. The same constraints allow identifiable parameters in our implementation. A distinction between our proposed framework and existing multivariate models is that we can potentially have only one observation at any of the spatial locations even when we have three response variables in our model. This makes it problematic to identify more than one latent process at each location. Our implementation using the SPDE approach and INLA avoids this problem since it does not directly model the latent variable parameters at the set of locations $\mathcal{U}$, but on the mesh vertices. When a model involves Matérn covariance function with a smoothness parameter greater than 1, the models using the SPDE approach can be used as an approximation while the model likelihood can be directly used with a Bayesian hierarchical approach.

The usage of our proposed framework is multifaceted. The interest sometimes lies within latent spatial processes when spatial data are analyzed, which represent residual spatial correlation in the response variables after considering existing covariates. The result can be used for detecting spatial clusters of unexplained risk or shared scientific drivers for response variables, which warrant further investigation in identifying those unknown drivers. When the interest is in predicting a response variable for a newly observed spatial unit, the fusion model improves the prediction of latent processes which in turn improves response variable prediction. Furthermore, the framework can be modified to use a one-dimensional Gaussian process in the latent components such that it applies beyond spatial data. For example, it can be used in time series modeling where all the observations are in $\mathcal{R}$ and as well as in machine learning applications (Rasmussen and Williams, 2005).

Although our current framework assumes independent latent Gaussian processes, it can be viewed as modeling multivariate Gaussian process via the LMC approach (Genton and Kleiber, 2015). One drawback is that the number of parameters in the design matrix $\boldsymbol{Z}$ increases with the number of latent Gaussian processes, which is difficult to estimate and may lead to convergence issues. In those cases, it may be advantageous to directly model the spatial structure with a cross-covariance function. In addition, the framework requires the number of latent processes q to be specified a priori.

As with any other statistical modeling, model misspecification has an impact on the inference and should be assessed via sensitivity analysis. Further research might include checking the compatibility of different data sources for spatial fusion modeling, i.e. if overlapping information exists between different spatial datasets. Such information helps to inform the model structure, especially the design matrix $\boldsymbol{Z}$.

Supporting information

Supplementary material for this article is available online at the author's website www.math.uzh.ch/furrer/download/unifying2020.zip, including all the R code used for the simulation studies and additional details on the Stan implementation of the framework using Bayesian hierarchical models.

Acknowledgments

We thank the Swiss Federal Statistical Office for providing mortality and census data and for the support which made the Swiss National Cohort and the application possible. The research was supported by the Swiss National Science Foundation (grants 175529, 3347CO-108806, 33CS30-134273 and 33CS30-148415). Members of the Swiss National Cohort Study Group are Matthias Egger (Chairman of the Executive Board), Adrian Spoerri and Marcel Zwahlen (all Bern), Milo Puhan (Chairman of the Scientific Board), Matthias Bopp (both Zurich), Martin Röösli (Basel), Murielle Bochud (Lausanne) and Michel Oris (Geneva).

References

Banerjee, S., Carlin, B.P., Gelfand, A.E., 2014. Hierarchical Modeling and Analysis for Spatial Data. CRC Press, pp. 136–139.
Banerjee, S., Gelfand, A.E., Finley, A.O., Sang, H., 2008. Gaussian predictive process models for large spatial data sets. J. R. Stat. Soc. Ser. B Stat. Methodol. 70 (4), 825–848.
Berrocal, V.J., Gelfand, A.E., Holland, D.M., 2010. A spatio-temporal downscaler for output from numerical models. J. Agric. Biol. Environ. Stat. 15 (2), 176–197.
Besag, J., York, J., Mollié, A., 1991. Bayesian image restoration, with two applications in spatial statistics. Ann. Inst. Statist. Math. 43 (1), 1–20.
Bopp, M., Spoerri, A., Zwahlen, M., Gutzwiller, F., Paccaud, F., Braun-Fahrländer, C., Rougemont, A., Egger, M., 2008. Cohort profile: The swiss national cohort — a longitudinal study of 6.8 million people. Int. J. Epidemiol. 38 (2), 379–384.
Chammartin, F., Probst-Hensch, N., Utzinger, J., Vounatsou, P., 2016. Mortality atlas of the main causes of death in Switzerland, 2008-2012. Swiss Med. Wkly. 146, 1–13.
Cressie, N., 1991. Statistics for Spatial Data. John Wiley & Sons.
Cressie, N., Johannesson, G., 2008. Fixed rank kriging for very large spatial data sets. J. R. Stat. Soc. Ser. B Stat. Methodol. 70 (1), 209–226.
Datta, A., Banerjee, S., Finley, A.O., Gelfand, A.E., 2016. Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets. J. Amer. Statist. Assoc. 111 (514), 800–812.
Diggle, P., Lophaven, S., 2006. Bayesian geostatistical design. Scand. J. Stat. 33 (1), 53–64.

Diggle, P.J., Moraga, P., Rowlingson, B., Taylor, B.M., et al., 2013. Spatial and spatio-temporal log-Gaussian Cox processes: extending the geostatistical paradigm. Statist. Sci. 28 (4), 542–563.
Follestad, T., Rue, H., 2003. Modelling Spatial Variation in Disease Risk using Gaussian Markov Random Field Proxies for Gaussian Random Fields. Technical report, Norwegian University of Science and Technology.
Fuentes, M., Raftery, A.E., 2005. Model evaluation and spatial interpolation by Bayesian combination of observations with outputs from numerical models. Biometrics 61 (1), 36–45.
Fuglstad, G.-A., Simpson, D., Lindgren, F., Rue, H., 2019. Constructing priors that penalize the complexity of Gaussian random fields. J. Amer. Statist. Assoc. 114 (525), 445–452.
Furrer, R., Genton, M.G., Nychka, D., 2006. Covariance tapering for interpolation of large spatial datasets. J. Comput. Graph. Statist. 15 (3), 502–523.
Gatrell, A.C., Bailey, T.C., Diggle, P.J., Rowlingson, B.S., 1996. Spatial point pattern analysis and its application in geographical epidemiology. Trans. Inst. Br. Geogr. 21 (1), 256–274.
Gelfand, A.E., Diggle, P., Guttorp, P., Fuentes, M., 2010. Handbook of Spatial Statistics. CRC press, p. 507.
Gelfand, A.E., Zhu, L., Carlin, B.P., 2001. On the change of support problem for spatio-temporal data. Biostatistics 2 (1), 31–45.
Genton, M.G., Kleiber, W., 2015. Cross-covariance functions for multivariate geostatistics. Statist. Sci. 30 (2), 147–163.
Geoinformation Kanton Zurich, 2015. PM10 - immissionen. https://opendata.swiss/de/dataset/pm10-immissionen1. Accessed: 2020–12–01.
Gotway, C.A., Young, L.J., 2002. Combining incompatible spatial data. J. Amer. Statist. Assoc. 97 (458), 632–648.
Greenland, S., 1992. Divergent biases in ecologic and individual-level studies. Stat. Med. 11 (9), 1209–1223.
Jin, X., Carlin, B.P., Banerjee, S., 2005. Generalized hierarchical multivariate CAR models for areal data. Biometrics 61 (4), 950–961.
Kelsall, J., Wakefield, J., 2002. Modeling spatial variation in disease risk: A geostatistical approach. J. Amer. Statist. Assoc. 97 (459), 692–701.
Knorr-Held, L., Best, N.G., 2001. A shared component model for detecting joint and selective clustering of two diseases. J. R. Stat. Soc. Ser. A 164 (1), 73–85.
Krainski, E., Gómez Rubio, V., Bakka, H., Lenzi, A., Castro-Camilo, D., Simpson, D., Lindgren, F., Rue, H., 2018. Advanced Spatial Modeling with Stochastic Partial Differential Equations Using R and INLA. Chapman and Hall/CRC.
Kyriakidis, P.C., Kim, J., Miller, N.L., 2001. Geostatistical mapping of precipitation from rain gauge data using atmospheric and terrain characteristics. J. Appl. Meteorol. 40 (11), 1855–1877.
Lindgren, F., Rue, H., 2015. Bayesian spatial modelling with R-INLA. J. Stat. Softw. Art. 63 (19), 1–25.
Lindgren, F., Rue, H., Lindström, J., 2011. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. J. R. Stat. Soc. Ser. B Stat. Methodol. 73 (4), 423–498.
Liu, Z., Le, N.D., Zidek, J.V., 2011. An empirical assessment of Bayesian melding for mapping ozone pollution. Environmetrics 22 (3), 340–353.
Lozano, R., et al., 2012. Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010. Lancet 380 (9859), 2095–2128.
Lunge Zürich, 2017. The LuftiBus project. http://www.lunge-zuerich.ch/de/projekte/luftibus/. Accessed: 2017–02–28.
MacNab, Y.C., 2016. Linear models of coregionalization for multivariate lattice data: a general framework for coregionalized multivariate CAR models. Stat. Med. 35 (21), 3827–3850.
Martins, T.G., Simpson, D., Lindgren, F., Rue, H., 2013. Bayesian computing with INLA: New features. Comput. Statist. Data Anal. 67, 68–83.
McMillan, N.J., Holland, D.M., Morara, M., Feng, J., 2010. Combining numerical model output and particulate data using Bayesian space–time modeling. Environmetrics 21 (1), 48–65.
Menezes, A.M.B., Pérez-Padilla, R., Wehrmeister, F.C., Lopez-Varela, M.V., Muiño, A., Valdivia, G., Lisboa, C., Jardim, J.R.B., de Oca Maria, M., Talamo, C., Bielemann, R., Gazzotti, M., Laurenti, R., Celli, B., Victora, C.G., for the PLATINO team, 2014. FEV_1 is a better predictor of mortality than FVC: The PLATINO Cohort Study. PLOS ONE 9 (10), 1–10.
Møller, J., Syversveen, A.R., Waagepetersen, R.P., 1998. Log Gaussian cox processes. Scand. J. Stat. 25 (3), 451–482.
Moraga, P., Cramb, S.M., Mengersen, K.L., Pagano, M., 2017. A geostatistical model for combined analysis of point-level and area-level data using INLA and SPDE. Spatial Stat. 21, 27–41.
Ogata, Y., 1988. Statistical models for earthquake occurrences and residual analysis for point processes. J. Amer. Statist. Assoc. 83 (401), 9–27.
R Core Team, 2018. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
Rasmussen, C.E., Williams, C.K.I., 2005. Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning). The MIT Press.
Ren, Q., Banerjee, S., 2013. Hierarchical factor models for large spatially misaligned data: a low-rank predictive process approach. Biometrics 69 (1), 19–30.
Rue, H., Held, L., 2005. Gaussian Markov Random Fields: Theory and Applications. CRC press.
Rue, H., Martino, S., Chopin, N., 2009. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. J. R. Stat. Soc. Ser. B Stat. Methodol. 71 (2), 319–392.
Sahu, S.K., Gelfand, A.E., Holland, D.M., 2010. Fusing point and areal level space-time data with application to wet deposition. J. R. Stat. Soc. Ser. C. Appl. Stat. 59 (1), 77–103.
Schmidt, A.M., Gelfand, A.E., 2003. A Bayesian coregionalization approach for multivariate pollutant data. J. Geophys. Res.: Atmos. 108 (D24), 1–9.
Shi, H., Kang, E.L., 2017. Spatial data fusion for large non-Gaussian remote sensing datasets. Stat 6 (1), 390–404.
Simpson, D., Illian, J.B., Lindgren, F., Sørbye, S.H., Rue, H., 2016. Going off grid: computationally efficient inference for log-Gaussian Cox processes. Biometrika 103 (1), 49–70.
Stein, M.L., 2008. A modeling approach for large spatial datasets. J. Korean Stat. Soc. 37 (1), 3–10.
Wackernagel, H., 2003. Multivariate Geostatistics: An Introduction with Applications. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 1–5.
Wang, C., Puhan, M.A., Furrer, R., 2018. Generalized spatial fusion model framework for joint analysis of point and areal data. Spatial Stat. 23, 72–90.
Wang, F., Wall, M.M., 2003. Generalized common spatial factor model. Biostatistics 4 (4), 569–582.
Wilson, K., Wakefield, J., 2018. Pointless spatial modeling. Biostatistics 1–16.
Zhang, L., Datta, A., Banerjee, S., 2019. Practical Bayesian modeling and inference for massive spatial data sets on modest computing environments. Stat. Anal. Data Min.: ASA Data Sci. J. 12 (3), 197–209.